Category Trend Analysis: Discovering the Top Relevant Trends in Your Data

Alina Maor, Renato Keshet

Hewlett Packard Labs
HPE-2016-33

Keyword(s):
Insights; Non-parametric; Ranking; Linear Regression; Scoring; Statistics; System; Time series analysis; Trend analysis; Weights

Abstract:
In this paper we present a novel approach for trend mining. Given a table of events, we generate all time series (signals) associated with all data sub-categories in the table and score each one according to their trend persistence, significance, discrepancy, alignment with other series, and user interest. For example, consider a table of product sales. Hundreds or thousands of time-series can be associated with each entity (e.g., vendors, products, etc.) in each possible market (e.g., geography, product category, etc.). For each such time-series, we compute its trend persistence (monotonic signals are scored higher), and its trend strength (higher scores for steeper slopes). In addition, we check the trend’s discrepancy to others in the same market (higher scores for relatively-high trends in low-activity markets), and compare between the signal trend and the market trend (scoring higher signals that change faster than their market). Finally, we score signals by the user interest, as expressed by user-defined weights. For trend-persistence calculation, we extend the Mann-Kendall test for trends to include arbitrary temporal weights. This allows one to implement time relaxation (“forgetting”), as well as include seasonality into the computation.
Category Trend Analysis:
Discovering the Top Relevant Trends in Your Data

Alina Maor and Renato Keshet
Hewlett Packard Labs,
Guthwirth Park, Technion, Haifa 32000, Israel
{alina.maor,renato.keshet}@hpe.com
http://www.hpe.com

Abstract. In this paper we present a novel approach for trend mining. Given a table of events, we generate all time series (signals) associated with all data sub-categories in the table and score each one according to their trend persistence, significance, discrepancy, alignment with other series, and user interest.

For example, consider a table of product sales. Hundreds or thousands of time-series can be associated with each entity (e.g., vendors, products, etc.) in each possible market (e.g., geography, product category, etc.). For each such time-series, we compute its trend persistence (monotonic signals are scored higher), and its trend strength (higher scores for steeper slopes). In addition, we check the trend’s discrepancy to others in the same market (higher scores for relatively-high trends in low-activity markets), and compare between the signal trend and the market trend (scoring higher signals that change faster than their market). Finally, we score signals by the user interest, as expressed by user-defined weights.

For trend-persistence calculation, we extend the Mann-Kendall test for trends to include arbitrary temporal weights. This allows one to implement time relaxation (“forgetting”), as well as include seasonality into the computation.

In order to demonstrate the approach, we have built a system that yields the top user-relevant trends from a given table of events. We have applied the system to the IDC Worldwide Quarterly Server Tracker report, and were able to spot the abrupt growth of Cisco in the market of "Blade X86" in North America, starting the third quarter of 2009, which was then overlooked by the competitors. The system provides significant insights for other sales data, as well as data from various managed areas, such as sociology, health, and IT operations.

Keywords: Non-parametric, Ranking, Linear regression, Scoring, Statistics, System, Time series analysis, Trend analysis, Weights

1 Introduction

“What is trending today?” This is the question that guides all major social networks: News outlets, Facebook, YouTube, Instagram and others. Trending
here means a consistent increase (or decrease) over time of a particular topic (usually measured by counting occurrence). Significant trends are important as they indicate where one might want to focus their attention and effort. This is not different in enterprise data analysis. Timely discovering the significant trends in a market or a category is key for successful decision making.

According to [1], trend analysis of time-series is aimed at allowing researchers and system users to build signal models (see [2] and references therein), and to use these models in order to predict the future signal samples (e.g., stock values [3], monthly water quality [4], or else). The deviations from predictions are typically used for evaluation of potential signal abnormality [5]. In this context, significant effort is placed on accurate data modeling [6]. For example, there are systems that predict which stocks are good for investment and investment automation [7]. This automation is often based on human defined strategies, stock-pool choice, risk policy and often demands close monitoring [7, 8].

A different task of trend analysis is signal monitoring for early abnormalities detection, preferably before a failure occurs. Often, this process involves automatic [9] or user-guided setting of rules or thresholds on signals behavior [10–12]. Signals are monitored in parallel, but each signal is analyzed independently from the others. The goal of such analysis is typically an alert (by e-mail, by alarm) to trigger human response (e.g., IT manager or security team intervention) or automated response (e.g. shutting down a machine) [5, 13, 14].

In this paper, it is not our goal to predict future samples, nor to fire predefined alerts. Our aim is to discover the significant trends in a dataset, among all possible combinations of players and categories. For instance, in a sales table, we would like to be able to find out that, just for illustration, sales in Brazil of product “laptops” in category “government agencies” has dropped significantly throughout the year. Given that there are almost 200 countries in the world, there may by several different buyer categories, with dozens or hundreds of products each, an analyst has to inspect hundreds to thousands of different country/product/category combinations. Time-series analysis when there are many time series to be monitored at once is an important topic in research, as well as real life applications. Examples of real-life fields with such multiple time-series are finance, complex machinery, security, and medical. Human analyst(s) is(are) physically incapable to monitor on-line (or even off-line) those signals, seeking significant trends. Therefore, automation of monitoring process is required.

For revenue management, every quarter, analysts process many thousands of markets with hundreds (or thousands) of players and decide which of the combination of these “conceal” new opportunities, emerging competitors, and decrease of markets. In this case, automatic statistical thresholds may not suffice. For example, a market with quarterly revenue of milliards that falls 2-sigma may be more relevant that a market with quarterly revenue of thousands (in whatever currency we measure the revenue). On the other hand, a player in a market of

\footnote{We will interchangeable use terms time-series and signals, often preferring the second term for its shortness.}
billiards that grew by 0.01 standard deviation may present more interest than a player in market of a million that raised 2-sigma of the market yearly growth.

Manual timely adaptation of rules for each market that accounts for all interplays and relations between different markets and market players is usually infeasible. Instead, there is a focus on modeling targeted forecasts based on additional knowledge, such as customer loyalty, products similarity, etc. [17, 18].

As analysts study the financial data of various markets and market players, they try to recognize anomalous events, e.g., emerging trends, and focus on these. A common tool for market analysis is visual inspection of financial charts (e.g., revenue). There are companies that are specializing on presenting the reports in a digestible form [19, 20]. But the time of the analysts is limited, so even when presented with a helpful visualization, they simply cannot cover large volumes of data and may miss important competitors. For instance, when Cisco rose in North America in the “Blade” server market in 2009 [21], they “won” a very large market share (roughly 40% of the revenue by 2014) [22] without competitors spotting it in time. Data reduction is sometimes done, e.g., by clustering data into segments, performing targeted forecasting and focusing only on large deviations from the expected. This type of analysis often results in missing valuable insights, as the important information may be “under the radar”.

In many datasets, (e.g., in finance, security, IT operations), there is much fluctuation in the time signals. Those signals that, on the other hand, maintain consistent trends are typically interesting, indicating a changed behavior in the data. These buried trends are what this work seeks to uncover. Our goal is to bring these significant trends to the user in a comprehensive form, i.e., explaining why a certain trend was scored higher than another. We present a method for signal trend calculation and corresponding scoring mechanism according to persistency and strength, consistency with respect to other signals in their group of reference, as well as the group dynamics itself, and by their “interest” to the user. We then return to the user the top trends according to the above score.

This work uses non-parametric trend detection technique combined with trends prioritization by user interests. The most common approaches for non-parametric trend detection are sign tests, rank tests or variations of these (see [23, 24] and references therein). Often, regressor fittings are used for this purpose too [25]. We use a count-based technique for trend detection which we enhance with trend slope estimation based on linear regression analysis. This is because we are looking for fast methods for approximating trend existence, persistence and strength rather than explicitly and accurately calculating it.

In the trend discovery stage, we compute trend significance weighted over time. I.e., we allow forgetting or emphasizing of different parts of the time-series. For example, we may give less significance to lack of trend in the distant past, but may as well benefit from existence of such a “long” trend-tail in prioritizing trend significance to the user. Another obvious example for weighted trends is seasonality weighting. For long trends, the computational complexity of trend identification does not scale up with the trend length, as there is a closed formula for trend significance which might be calculated off-line for each trend-length.
We begin with a weighted sign-changes based approach, enriching it with linear regression-approximation of the trend. We determine the most persistent (stable) trends, provide their ranking according to the “anomaly” criterion and present the top-most results to the system user along with a score reasoning.

Trend discovery is based on the non-parametric algorithm for trend detection – the Mann-Kendall test([26, 27]). In the next two sections we overview the Mann-Kendall algorithm (Section 2) and demonstrate how we change it to include weights (Section 3) which are not necessarily seasonal as in [28]. In Section 3, we then introduce a formula for the trend score and discuss the importance of data preprocessing. In Section 4, we conclude with describing our system and showing the results of the trends scoring applied to IDC Worldwide Quarterly Server Tracker, Q1, 2015, report [29].

2 Mann-Kendall trend test

One of the most known methods for trend analysis is what is called the Mann-Kendall test [26], [27]: Consider times between \( t_2 \) and \( t_1 \), when \( t_1 < t_2 \), for a given time-series \( \{X_t\} \). If there is no trend in this time-interval and \( \{X_t\} \) is i.i.d., the sum of signs of changes between all possible between-sample differences should be around zero. The Mann-Kendall test checks existence of trend against this hypothesis. If the hypothesis is rejected, it means that there is a trend.

Given a time-series \( \{X_t\} \) of length \( n \), we assume that the null hypothesis of data randomness \( H_0 \) is that the time series are a sample of i.i.d. random variables. The alternative hypothesis, \( H_1 \), of a two-sided test is that the distribution of the series is not identical for all samples. The test statistics \( S \) [28] is defined as:

\[
S = \sum_{k=1}^{n-1} \sum_{j=k+1}^{n} \text{sgn}(X_j - X_k) \tag{1}
\]

where \( n \) is the length of the given time-series and \( \text{sgn}(\theta) \) is a sign function [30]. In [26], it had been shown that under \( H_0 \), the \( S \) is normally distributed as \( n \) reaches infinity, with zero mean, \( E[S] = 0 \). Mann has treated the problem assuming that there are no ties in the time-series and derived an explicit expression for variance of \( S \) under these conditions. Kendall (e.g., [27]) had further developed the expression of \( \text{Var}[S] \) under the possibility of existence of ties, obtaining:

\[
\text{Var}[S] = \frac{n(n-1)(2n+5)}{18} - \sum_k t(t-1)(2t+5), \tag{2}
\]

where \( t \) is the length of a tie (number of identical signal values) and \( k \) is the number of ties in signal. Notice, though, that if there are ties in the data, their influence on \( S \) will be zero and hence by taking a larger variance (that includes ties) we only harden the trend-test. Both Mann and Kendall showed that for \( n \) above 10 the normal approximation of the distribution of \( S \) is excellent. Hence, for \( n > 10 \) “it is correct” to use the normal approximation, computing \( Z \) random
variable, as given next, and perform hypothesis testing on it:

\[ Z = \begin{cases} \frac{S-1}{\sqrt{\text{Var}[S]}} & \text{if } S > 0, \\ 0 & \text{if } S = 0, \\ \frac{S+1}{\sqrt{\text{Var}[S]}} & \text{if } S < 0. \end{cases} \quad (3) \]

In a two-sided test for trend, \( H_0 \) is accepted if \(|Z| \leq z_{\alpha/2}\), where \( F_N(z_{\alpha/2}) = \alpha/2 \), \( F_N \) being the standard normal cumulative distribution function and \( \alpha \) is the size of the significance level for the test. The sign of \( S \) (or \( S - 1 \)) indicates if the trend is positive or negative. The shift of \(+1\) in (3) came as a hardening tool for the hypothesis testing [27], but it is not compulsory. For low orders of \( n \), \( n \leq 10 \), critical values are given using combinatorics.

The authors of [28] addressed existence of seasonality. In [28], it was proposed to first divide the signal into seasons, to compute \( S_{s_i} \) per each season \( s_i \), and to use the sum of \( S_{s_i} \) as \( S \) in (3) and to proceed with hypothesis testing.

This approach actually raises a broader question — what if not all data samples are equally important? Suppose that we want to give higher weights to recent samples “forgetting” samples from distant past? And suppose that we also want to incorporate seasonality into this framework. How would we change the test to address these points? We answer this question in the next section.

There are many extensions and applications of the Mann-Kendall test. In part, [32] is about removal of correlation between signal values which may bias the test. We do not remove correlation as we only want an approximation of trend existence, but, we use insights from [32] to include weights into the test.

3 Category Trend Analysis

3.1 Trend test with temporal weights — long trends

Our objective is to determine if there is a trend in the observed signal. Trend testing is done in an \( n \)-sample window. But what if we do not want to consider only \( n \) samples, but we want to use historical data? We then wish to give more influence to the recent samples. Yet, we do not want to omit the influence of old values in case that they support the trend conjecture! We merely want to tone these down. Seasonality cropping is another example where we weight samples. Hence, we propose to use a modified trend-test. The results of [28] are then become a private case of the problem that we solve as we show in the following.

The definition of \( S \) in (1) is changed as follows:

\[
S_w = \sum_{k=1}^{t-1} \sum_{j=k+1}^{t} W_{i,j} \text{sgn}(X_j - X_k), \quad (4)
\]

where \( \{W_{i,j}\} \) is the set of all temporal weights to be applied between time \( i \) and time \( j \). For example, if one is interested in suppressing of influence of past samples, the weight \( W_{i,j} = W_j \) will become dependent on index \( j \) only, reducing
the weights as $j$ becomes smaller, for example with exponential decay. If one wants to take into account relative importance of these samples with respect to other samples, e.g., accounting for seasonality as in [28], we set weights between seasons to 0 while giving weights of 1 to samples belonging to the same season.

Let us substitute the sign function in (1) with definition from [32]:

$$a_{i,j} = \text{sgn}(x_j - x_i) = \text{sgn}(R(x_j) - R(x_i)).$$

(5)

Now, we use the following observation from [32]: The variance of $S$ does not depend on the distribution of $X$. This is because $a_{i,j}$ that build (1) can be expressed via the ranks of samples rather than their particular values. This is true as eq. (5) is defined by ordered differences of $x_j$ and $x_i$, so in principle, this difference only depends on rank of $x$.

Following [32], if $X$ is $N(\mu, \sigma)$, then the difference $(x_j - x_i) \sim N(0, 2\sigma)$. Therefore, as a sum of normal variables is normal, we already know that $S_w$ denoted in (4) is normal too. It’s left to find the mean and the variance of $S$:

$$E(S_w) = E \left[ \sum_{k=1}^{t-1} \sum_{j=k+1}^{t} W_{i,j} \text{sgn}(X_j - X_k) \right]$$

(6)

$$= \sum_{k=1}^{t-1} \sum_{j=k+1}^{t} W_{i,j} E[\text{sgn}(X_j - X_k)]$$

(7)

$$= 0,$$

(8)

so, $E(S_w)$ has not changed by adding weights. As for the variance of $S_w$,

$$\text{Var}(S_w) = E(S_w - E(S_w))^2 = E(S_w^2),$$

(9)

where the last equality follows from (8). Following [32],

$$\text{Var}(S_w) = E \left[ \sum_{k=1}^{n-1} \sum_{j=k+1}^{n} W_{k,j} \text{sgn}(X_j - X_k) \right]^2$$

(10)

$$= \sum_{k=1}^{n-1} \sum_{j=k+1}^{n} W_{k,j}^2 E[a_{k,j}^2]$$

(11)

$$+ \sum_{k=1}^{n-1} \sum_{j=k+1}^{n} \sum_{i=k+1}^{n} W_{k,j} W_{k,i} E[a_{k,j} a_{k,i}]$$

(12)

$$+ \sum_{i=1}^{n-1} \sum_{j=k+1}^{n} \sum_{k \neq i, k \neq j, l=k+1}^{n-1} W_{i,j} W_{k,l} E[a_{i,j} a_{k,l}].$$

(13)
Now, as \( \{x_i\} \) are independent, \( E[a_{i,j}a_{k,l}] = 0 \) when \( i \neq j \neq k \neq l \), eliminating the third term in (10).

\[
E[a_{i,j}^2] = E[\text{sgn}(x_j - x_i)]^2 = \Pr(x_j \neq x_i).
\]

(14)

If we assume here that there are no ties in the series, the square of \( \text{sgn} \) is always 1 and \( E[a_{i,j}^2] = 1 \). Using the development of either Kendall or [32], when there are no ties, \( E[a_{i,j}a_{k,l}] = \frac{1}{n} \). Hence, without ties, (4) turns into

\[
\text{Var}(S_w) = \sum_{k=1}^{n-1} \sum_{j=k+1}^{n} W_{k,j}^2
\]

(15)

\[
+ \frac{2}{9} \sum_{k=1}^{n-1} \sum_{j=k+1}^{n} \left[ \sum_{i=j+1}^{n} [W_{k,j}W_{k,i} + W_{k,j}W_{j,i}] + \sum_{i=k+1}^{j} W_{k,j}W_{i,j} \right].
\]

Now we may proceed as previously for testing the hypothesis of having a trend in the system. If there are ties in the system, there are actually two approaches – one means that the significance of the test is actually less harsh than we expect (as we allow rather large variance for change detection), but this would also mean that had we determined a trend it’s a stronger trend. So we may leave the test as it is, declaring trends of high significance only. Alternatively, we may modify \( \text{Var}(S_w) \) to account for all possible ties by reducing their impact from (16) i.e., for each tie of length \( t \), removing the interplay between these places by subtracting from (16) for each tie \( t_k: \sum_{k=1}^{n} \sum_{j=k+1}^{n} W_{i,j} + \frac{2}{9} \sum_{k=1}^{n} \sum_{j=k+1}^{n} \left[ \sum_{i=j+1}^{n} [W_{k,j}W_{k,i} + W_{k,j}W_{j,i}] + \sum_{i=k+1}^{j} W_{k,j}W_{i,j} \right] \)

which results in:

\[
\text{Var}(S_w) = \sum_{k=1}^{n-1} \sum_{j=k+1}^{n} W_{k,j}^2
\]

\[
+ \frac{2}{9} \sum_{k=1}^{n-1} \sum_{j=k+1}^{n} \left[ \sum_{i=j+1}^{n} [W_{k,j}W_{k,i} + W_{k,j}W_{j,i}] + \sum_{i=k+1}^{j} W_{k,j}W_{i,j} \right]
\]

\[
- \sum_{k_1 \in \{k_i\}} \sum_{j_1 \in \{j_i\}} \sum_{k \in \{k_i\}} \sum_{j \in \{j_i\}} [W_{k,j}W_{k,i} + W_{k,j}W_{j,i}]
\]

\[
+ \frac{2}{9} \sum_{k_1 \in \{k_i\}} \sum_{j_1 \in \{j_i\}} \sum_{k_2 \in \{k_i\}} \sum_{j_2 \in \{j_i\}} [W_{k,j}W_{k,i} + W_{k,j}W_{j,i}]
\]

\[
+ \sum_{i \in \{k_i\}} W_{k,j}W_{i,j} \]

(16)

Note that for every value of \( n \), ties list \( \{k_i\} \) and given weights, it is possible to a-priori calculate the value in (16), store this in a look-up-table and use the desired one upon need. For equal weights, the test gets back to the original test and for that case we may use the function of \( \text{Var}(S_w) \) as a function of \( n \).
3.2 Trend test with temporal weights – short trends

We still need to address the issue of short trends, e.g., trends that are shorter than a certain \( n \), \((n = 9\) for the original Mann-Kendall test [27] or \( \{n_w\} = 3\) per season [28]). We compute explicitly the weighted probabilities for each given set of weights by noting that, for every given \( n \) we know exactly how many “types” of infinitely many vectors of distinct values there are ([26] or simple combinatorics):

There are only \( n! \) permutations of a given vector with \( n \) distinguishable values, hence we generate a set of such permutations for each \( n \) and calculate the explicit values of \( S_w \) for each given set of weights. Now we may obtain \( p \)-values from this “enriched” by weights but still symmetric distribution (there are more \( p \)-values then in the original Kendall-Mann version as weighting provides a higher variability of possible \( S_w \)-values). For each input vector we may calculate \( S_w \) and compare these to the obtained \( p \)-value for a given critical values \( \alpha \).

In practice, we will not work with very long trends (consistency over a year, i.e., 4 quarters, is already very meaningful), hence we do not perform too many calculations. The length of \( n \) for which we may use the long-trend formula varies for different sets of weights. Every particular choice of weights needs to be tested against validity of Gaussian approximation using standard techniques, e.g., as in [28]. When using equal weights, we get back to the formula developed in [31].

Moreover, as we are seeking approximation of the trend existence, in cases where there is a crucial need for speedy computations, we may, in fact, use the inaccurate Gaussian approximation from section 3.1 developed for large \( n \), for small values too, as it will still provide us with some (maybe inaccurate, but still hinting about the existing of the trend) results. In practical trials, for \( n \) as small as 5, we obtained sufficiently good (for our purpose) results when omitting ties from the equation formula, i.e., when we harden the trend test.

3.3 Trend mining – communicating trend insights to the user

Recall the toy-example of the finance use-case from the introduction – the analyst investigates many markets and players in markets and needs to decide which of the time series is “financially interesting”. We want to help the analyst by presenting the most significant trends.

The analysis is, of course, not constrained to financial markets and players and in the sequel we may as well assume that we have groups (markets) of comparable entities (players) which may or may not have trends, and we wish to rank the importance of showing a certain activity to the system user (monitoring personae, analyst or even a doctor conducting a clinical trial and monitoring many different patients. In the last case, the patient blood tests of different patients are the “players” and different blood types are the “markets”).

Trend test with the modified Mann-Kendall weighted test is, hence, only a brick in our journey to solving the general problem – we want to present the system user with the ability to investigate and prioritize system signals, according to their trend significance and interest for the user. For this we need to define a measure of interest of a signal.
In systems with inconsistent trends, consistent trends are of interest to the analyst, hence we would want to reflect this in our measure of interest. But it is not enough to bring the most persistent trends. It is also interesting to consider the impact of the trend on its group, its discrepancy from other group players and the change of the player relative to its group and across other groups.

We propose the following framework for bringing up insightful trends via their trend score, $T_s(\cdot, \cdot)$, based on three key parameters:

1. Trend stability/persistence significance - $T_p(\cdot, \cdot)$;
2. Trend discrepancy - $T_d(\cdot, \cdot)$;
3. Trend value to the user - $T_v(\cdot, \cdot)$.

Carrying on the toy example, let us uniquely identify each player, $p$ associated with a certain market, $m$, (the same player may belong to the same market – for example, product “Blade” may appear both in market {Vendor “HP”} and in market {Country = USA}). We define the trend score $T_s(p, m)$ for player (entity) $p$ in market (group) $m$ as a function of four conceptual variables:

$$T_s(p, m) = f(T_p(p, m), T_d(p, m), T_v(p, m), \bar{W}(p, m)).$$

where $\bar{W}$ are some additional, possibly external weights parameters (e.g., the user might decide to eliminate big markets or insignificant trends from presentation).

The trend stability$^2$ is determined by the trend significance value $Z$ (eq. (3)):

$$T_p(p, m) = |Z(p, m)|,$$

where $Z(p, m)$ is the weighted trend strength $S_w(p, m)$ scaled by the volatility given in eq. (16). Note the following – if we harden the trend test and omit ties, the volatility in (eq. (16)) is the same for all time series of a given length $n$. Hence, it may be omitted from the formula, leaving $T_p(p, m) = S_w(p, m)$.

The discrepancy of a certain trend, $T_d$, is a tricky thing to determine. We propose to use the following conceptual components in its determination:

1. Discrepancy of the trend significance of the entity $p$ within its group $m$,
2. Discrepancy of the trend change within the group,
3. and percentage discrepancy of the trend change within the group.

Let us make the following notation:

1. Slope of group $m$ - $L(m)$;
2. Percentage slope of group $m$ - $L_p(m)$;
3. Slope of entity $p$ - $L(p)$;
4. Percentage slope of entity $p$ - $L_p(p)$;
5. Standard deviation of absolute trend significance of all entities in group $m$ - $\Delta_s(\{p\} \in m|m)$;

$^2$ We extend on of appropriate affiliation of trend significance to player and market in the next subsection 3.4.
6. **Standard deviation of absolute trend slope of all entities in group** \( m \) \( \Delta_{st}(\{p\} \in m|m) \);  
7. **Group value** (e.g., size or revenue) for group \( m \) \( v_{g}(m) \);  
8. **Entity value** (e.g., size or revenue) for entity \( p \) in market \( m \) \( v_{e}(p, m) \);  
9. **Trend test output for entity** \( p \) in group \( m \) \( I(m; p) \)  

Here, \( I(m; p) = sgn(\cdot) \) is the sign function representing the output of the modified Mann-Kendall test. Hence, we omit signals with insufficiently significant trends from future analysis as their score is always zero. But we do use these trends \( Z \)-values to estimate relative statistics of entities and groups. Upon application, it is possible not to use this “hard decision” on trend significance, as the values of \( Z \) are smaller than these of significant trends “anyway”, but we prefer to omit all “insignificant” trends from future consideration as insignificant (weak or non-existing) trends are not interesting in our score measure.

In its simplified form, \( T_d \) is defined as following:

\[
T_d(p, m) = I(m, p) \cdot \left( \frac{T_p(p, m)}{\Delta_t(\{p\} \in m|m)} \right)^\alpha \cdot \left( \frac{|L(p)|}{\Delta_t(\{p\} \in m|m)} \right)^\beta \cdot \left( \frac{|L(p)|}{L(m)} \right)^\gamma \cdot \left( \frac{L_p(p)}{L_p(m)} \right)^\delta \quad (19)
\]

Note that we do not use the Mahalanobis distance (e.g., \( \frac{T_p(p, m) - \mu_s(\{p\} \in m|m)}{\Delta_t(\{p\} \in m|m)} \)) in definition of eq. (19) as we are not interested in relative distances but in absolute size of changes. Using the Mahalanobis distance would reduce the significance of entity changes that we measure by standard deviations.

Finally, we define the value to the user by the relevant value in the system which is in its simplified form is the size of the entity and the size of its group:

\[
T_v(p, m) = (v_e(p, m))^\lambda \cdot (v_g(m))^\omega \quad (20)
\]

This value is especially fitting the case of financial data, as the user is not interested in insights that do not “worth” money. It is always possible to alter this definition by providing external weights \( \tilde{W} \).

Here, the constants \( \alpha, \beta, \gamma, \delta, \lambda \) and \( \omega \) are user defined (either directly or via user interaction) and are preset on specific initial values, e.g., \( \alpha = 1, \beta = 0.5, \gamma = 0.5, \delta = 0.1 \) and \( \lambda = 0.8, \omega = 1.0 \).

The **most interesting time series** are these with highest \( |T_d(p, m)| \), where \(|\cdot|\) stands for absolute value. The sign of these \( |T_d(p, m)|, I(m, p) \), determines whether the trends are rising or declining.

We approximate the trend change by empirical linear regression [33]. Note that we give two possibilities of the slope interest – one is direct use of the result of the linear regression, while the other one the relative percentile change with respect to the starting point returned by the regression. The later definition gives additional measure to the significance of the trend change rather than measuring rate change. E.g., change in a big player that already starts from a big value may be somewhat less interesting than great percentile change in a small emerging competitor. Importance is set by the user via setting \( \gamma \) and \( \delta \).
3.4 Data preparation nuance

We pause to emphasize the following aspect of our analysis, which was directly inspired by financial problems, but which is, in fact, very general and strong—it is difficult to compare different entities and different groups trend-significance because the group sizes and the entity activity may be measured in “incomparable” quantities—some markets worth billions of dollars, while others merely scratch a million or less. The same goes for blood tests, where a normal range of one type of tests may be measured in milli-units and another in micro-units and almost any another application that we may keep in mind.

The trends associated to each of the entities are also relative to their group behavior in case of econometrics as well as for security and medical care. But if we transform the measurements relative to baseline (group) behavior (trend), it will be possible to compare between trends in relative percentiles rather than directly to work on given time-series. Hence (upon application), we propose to transform the original time-series into relative “market-share”-like percentage time-series [34] and apply trend-analysis there, while using both market-share and original time-series values in determination of the trend score elements.

In part, while we propose to keep most of the variables we defined in the previous section in their native units, we propose to calculate $S_w$ in the market-share domain, a transformation which is common in finance. Assume that $X_t(p, m)$ is the value of the signal $X$ at time $t$ describing entity $p$ in group $m$. Transformation of $X_t(p, m)$ into the percentage relative domain is [34]:

$$\hat{X}_t(p, m) = \frac{X_t(p, m)}{\sum_{i \in [p] \in m} X_t(i, m)}.$$  \hfill (21)

4 Results

4.1 Data-set description

In this example we use a financial IDC Tracker report[29]. This report accumulates various market and player statistics, such as the number of sold units for each type or family of products and the quarterly revenues in millions of dollars for a very large collection of vendors in a certain field of operation. In fact, the report provides revenues for a variety of parameters for each vendor, per geography, product family and product specifics. From this report it is possible to aggregate all possible combinations of fields describing each of the revenues into super-groups, or larger markets as well as examine players which are not directly indicated in the report, e.g., accumulate revenue values in Germany, over the 4-core products of family “Blades” for each of the vendors. This is the

---

[3] If there is a global change in network activity due to some new application affecting all computers in the network, the entire network flow will change too. Similarly, if a certain medication alters blood tests, the control-group test-baseline will change. In both cases if the change is dramatic and persistent, it will be spotted during evaluation of the baseline activity trend.
type of markets and players that we are referring to, in this example, vendor being the player in the above market.

4.2 System of Trend Analysis

We have built a system for trend mining using the concepts of Section 3. A system view is presented in Fig. 1: In the options field, the user is able to choose the period of interest. He or she may then choose the type of the revenue to be used – it may be either the value of the revenue itself or the number of units sold by the player in that category. The attributes are the features of the IDC report and the user may decide if to choose analysis of all markets at once or to choose particular sub-markets. In the case of the example, the user has chosen to focus on all markets which are subsets of “Vendor” and “Product” choice and the players are all instances of attributes that were clicked but was not included in the market description. An example of the Market-Player set is Market = “Vendor” and Player = “Product”, i.e., all product instances, which are in this case “Rack-Optimized”, “Blade”, “Tower” and “Density Optimized” are tested for trend over all Vendors (e.g., “IBM”, “AB”) competing against trends of Market = “Product” and so on. Note that we compare between all trends over all different markets, including the “global market”, where the players are all sub-sets of the chosen attributes.

The output of the system are the most significant (according to our notion of significance score) trends. These appear ranked in decreasing order of significance, in the top-most part of the screen, in code-colored boxes. Here, red indicates downward trends, green stands for up-rising trends and the color intensity is the overall trend significance. In grayscale, the intensity of the box is its significance. The graph is presented in Market Share rather than real revenues. For each meaningful trend, we provide the user with a chart of all other players in the market. It is possible, of course, to plot exact input values (revenues), but we could not show the numbers due to data confidentiality.

Recall the Cisco example we gave in the Introduction section. In 2010, the growth of Cisco was widely noticed both in the Blade and the global market in North America. But the question is, was it possible to spot this constant growth earlier? Our answer is affirmative. Specifically, we investigate early revenues between 2008Q3 and 2009Q3. In our toy example, we have not even constrained ourselves to any specific geography, as the growth was so big that it should have reflected globally. The results of the investigations are presented in Figs. 1: Cisco (the bold line) came as the 12th most significant trend among all players in the global market and 18th in the “Blade” market between the last quarters of 2008 and 2009! (The first, most significant trend is labeled with 0 index.) While it scored high, when one looks at the report market-share/revenue charts, it is not at all perceptible in the global market and is not even that visible in the “Blade” market, hence, spotting its growth based on charts only is nearly impossible.

The growth of Cisco became notable in 2010, stating growth since the third quarter of 2009. To zoom into 2010, consider Fig. 2 demonstrating a quarter over quarter comparison between the 2nd quarter of 2009 (Cisco hasn’t started
Fig. 1. Example: Cisco is the 12th “significant” trend between the last quarters of 2008 and 2009 in the global market and 18th “significant” trend in the “Blade” market.

its growth yet) and the 2nd quarter of 2010 – Cisco comes as the 4th significant trend in the global market and the 8th trend in the “Blade” market! Note that the chart already shows an “insignificant” trend of Cisco in the “Blade” market.

Due to shortness of space, we will not get into showing this vendor evolution in other markets or other time periods, but we believe that the point is clear. This example also demonstrates the power and flexibility of our measure – by tuning the importance of its components, it is possible to take generate different significance scores: For example, here, the size of the global market is larger than the size of the “Blade” market, hence the global market insight was scored higher. By reducing the impact of the market size we could make greater emphasis on, for example, long-term stability of the trend rather than emerging trend. Due to the lack of space, we do not show the explanatory visualization behind the insight, but it is obvious in this case – the raise of Cisco was constant, growth faster than the market and other competitors as well as being part of a large market. Here we used weak forgetting as the observation period is very short, but we obtained similar results considering longer weighted periods.

5 Conclusion

We have built a system that brings insights on time series activity and is able to highlight the most significant series for immediate user attention. We have
Fig. 2. Example: Cisco is the 4th “significant” trend between the second quarters of 2009 and 2010 in the global market and 8th “significant” trend in the “Blade” market.

developed a measure of trend-based significance for signal comparison and extended a non-parametric trend test to include forgetting of historical or simply irrelevant times from trend definition.

References


